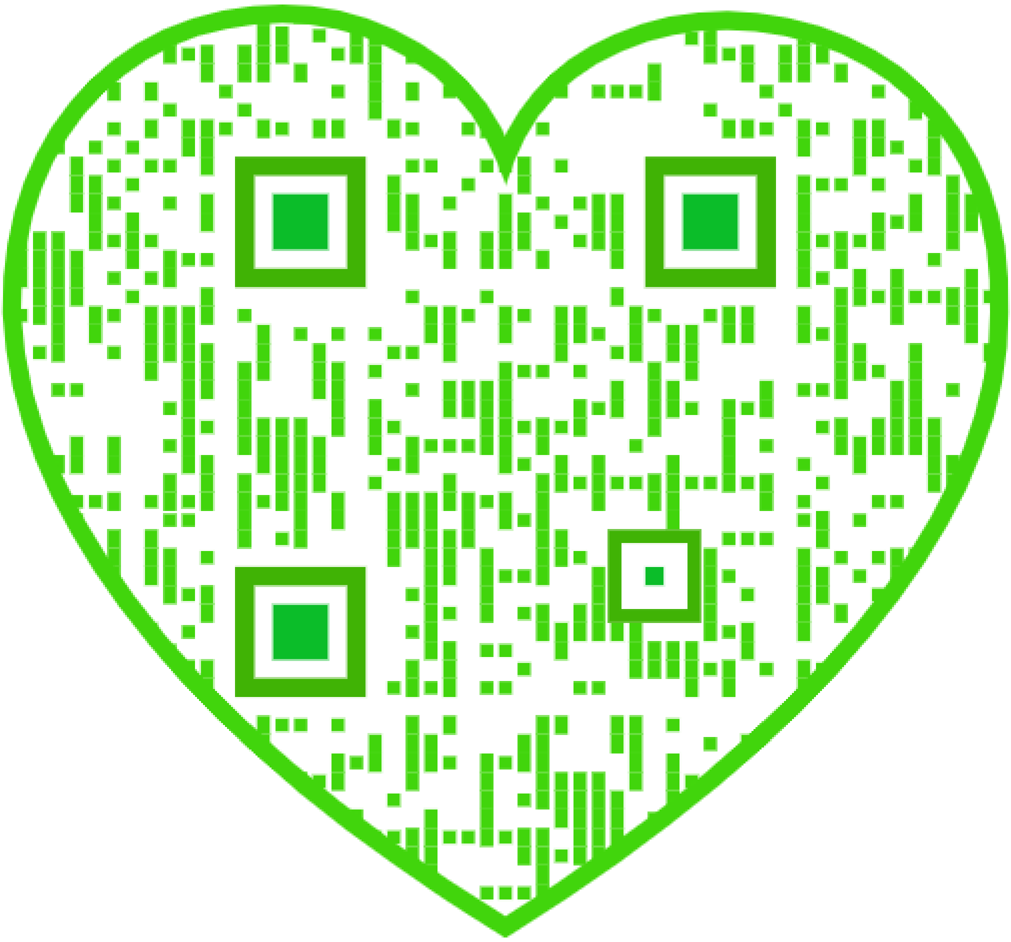
Master in Artificial Intelligence

Feature Engineering III



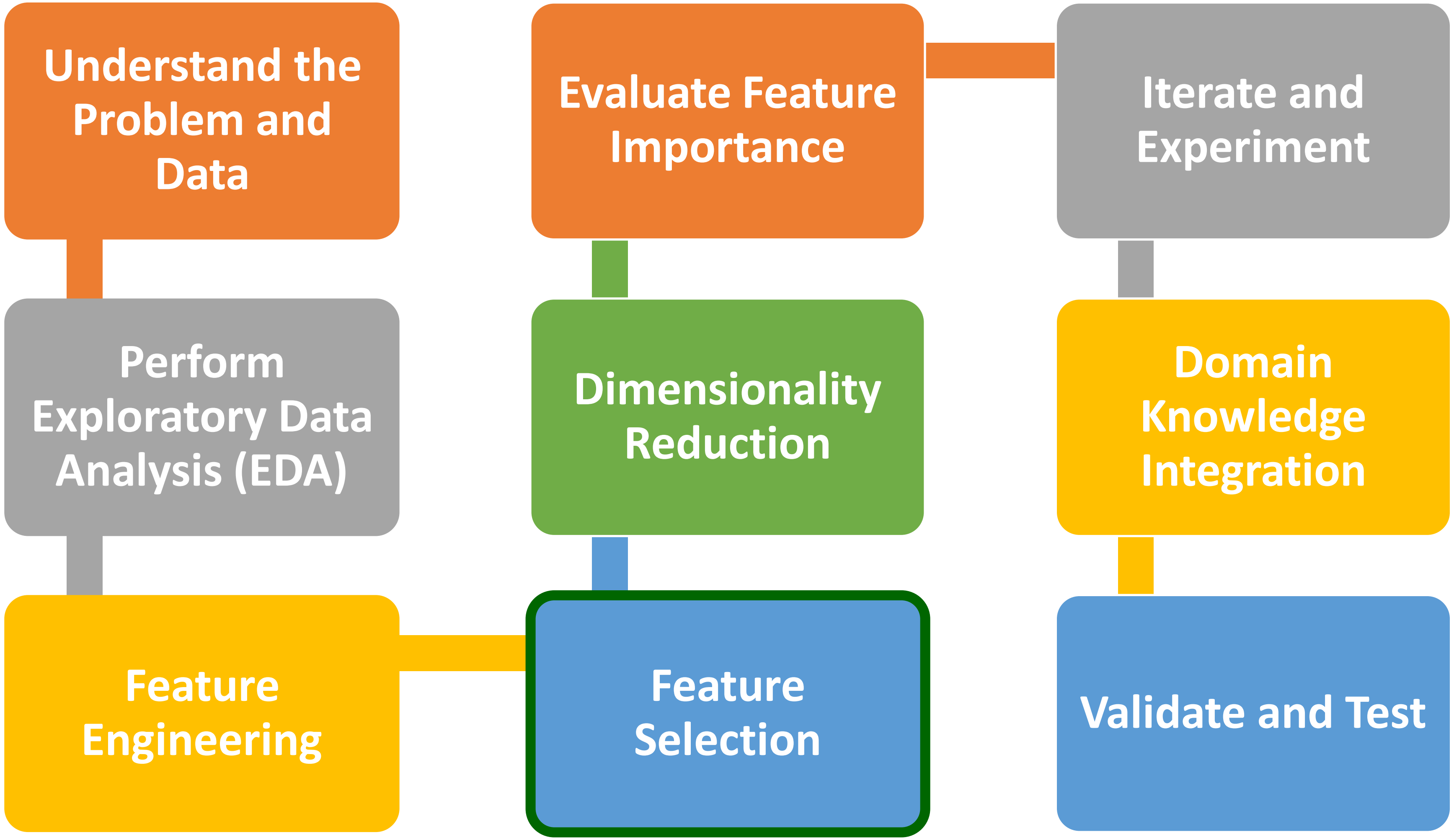
Purpose

The purpose of the section is to help you learn how to identify and extract meaningful features from the data to become a Successful Artificial Intelligence (AI) Engineer

At the end of this lecture, you will learn the following

How to select a subset of the most relevant features

How to select a subset of the most relevant features



Understand the Problem and Data

Perform Exploratory Data Analysis (EDA)

Feature Engineering

Feature Selection

Dimensionality Reduction

Evaluate Feature Importance

Iterate and Experiment

Domain Knowledge Integration

Validate and Test

How to select a subset of the most relevant features

Univariate Feature Selection

- Based on statistical tests such as chi-square, ANOVA, or mutual information

Recursive Feature Elimination (RFE)

- Iteratively remove least important features based on model performance

Feature Importance

- Assess the importance of features using ensemble methods like Random Forests or Gradient Boosting

How to select based on statistical tests

Univariate Feature Selection

- Based on statistical tests such as chi-square, ANOVA, or mutual information

Recursive Feature Elimination (RFE)

- Iteratively remove least important features based on model performance

Feature Importance

- Assess the importance of features using ensemble methods like Random Forests or Gradient Boosting

How to select based on Chi-Square Test

The chi-square test is used to assess the independence between categorical variables. It measures the significance of the association between a categorical feature and a categorical target variable

```
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2

# Assuming X contains your features and y contains your target variable
# Select the k best features based on the chi-square test
k = 5  # Number of features to select
selector = SelectKBest(score_func=chi2, k=k)
X_new = selector.fit_transform(X, y)
```

How to select based on ANOVA (Analysis of Variance)

ANOVA is used to assess the significance of differences in the means of numerical features across different categories of a categorical target variable

```
from sklearn.feature_selection import f_classif

# Assuming X contains your features and y contains your target variable
# Compute ANOVA F-values and select the best features
selector = SelectKBest(score_func=f_classif, k=k)
X_new = selector.fit_transform(X, y)
```

How to select based on Mutual Information

Mutual information measures the dependency between two variables, regardless of their types (categorical or numerical). It quantifies the amount of information obtained about one variable through the other

```
from sklearn.feature_selection import mutual_info_classif

# Assuming X contains your features and y contains your target variable
# Compute mutual information scores and select the best features
selector = SelectKBest(score_func=mutual_info_classif, k=k)
X_new = selector.fit_transform(X, y)
```

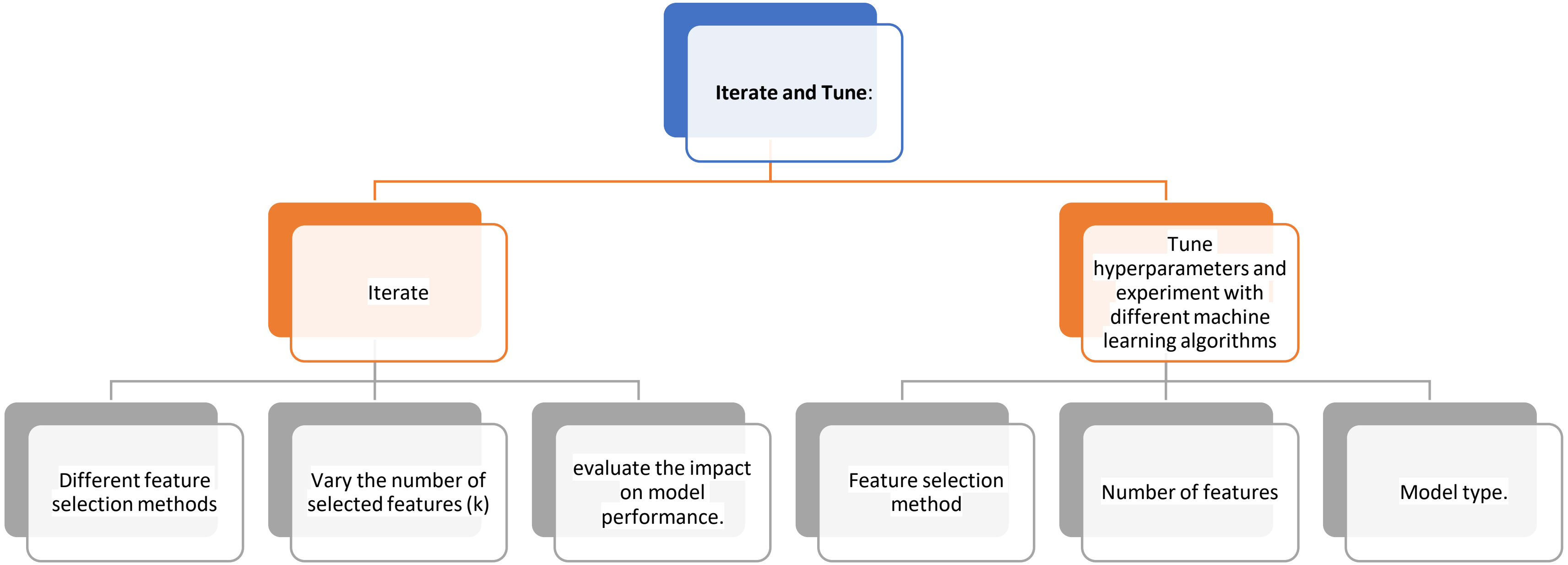
How to select based on statistical tests

```
from sklearn.model_selection import cross_val_score
from sklearn.linear_model import LogisticRegression

# Assuming X_new contains the selected features
# Initialize your model
model = LogisticRegression()

# Evaluate model performance using cross-validation
scores = cross_val_score(model, X_new, y, cv=5)  # Adjust cv as needed
print("Mean Accuracy:", scores.mean())
```

How to select based on statistical tests



Iterate and Tune:

Iterate

Tune hyperparameters and experiment with different machine learning algorithms

Different feature selection methods

Vary the number of selected features (k)

evaluate the impact on model performance.

Feature selection method

Number of features

Model type.

How to select a subset of the most relevant features

Univariate Feature Selection

- Based on statistical tests such as chi-square, ANOVA, or mutual information.

Recursive Feature Elimination (RFE)

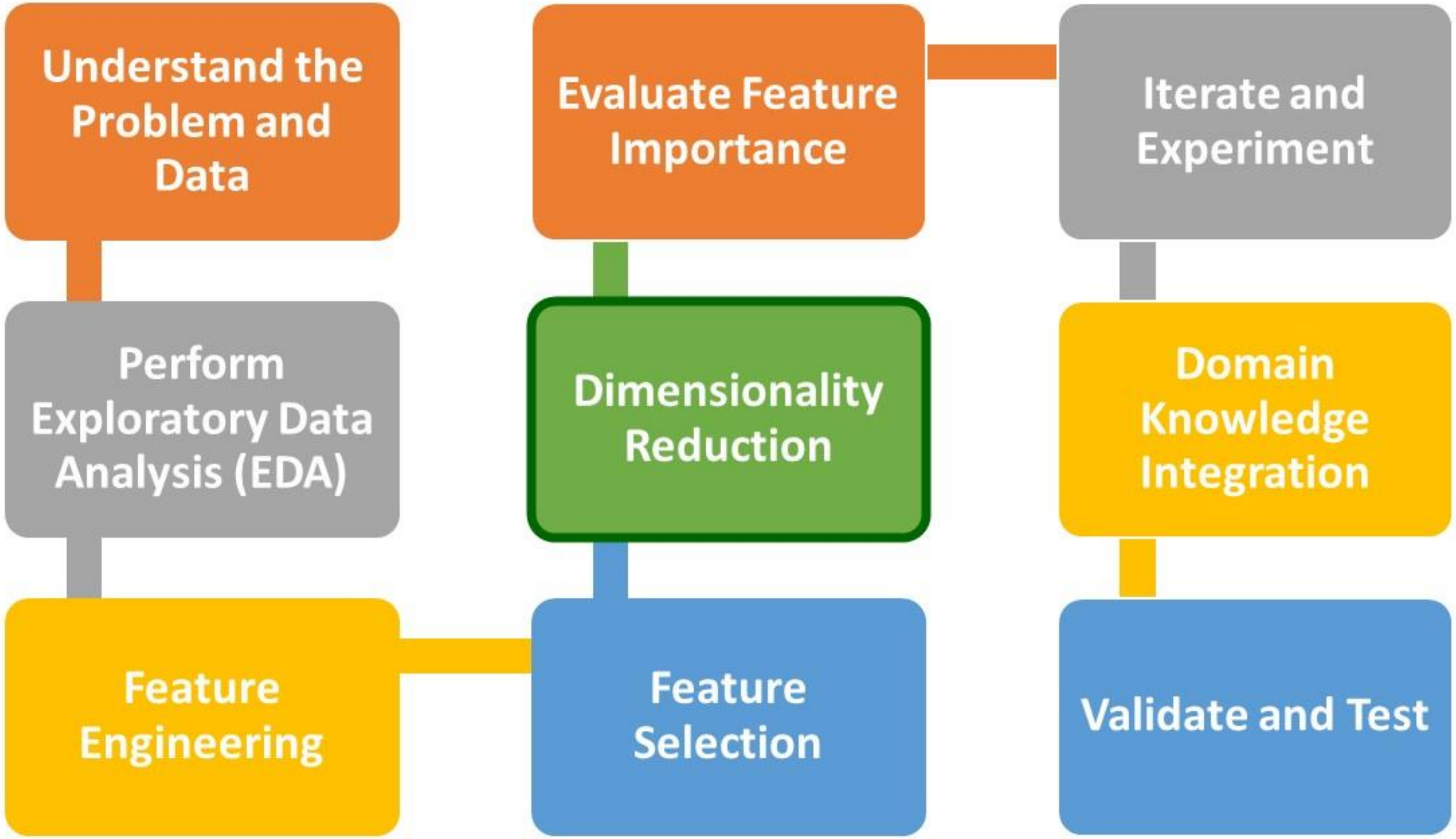
- Iteratively remove least important features based on model performance.

Feature Importance

- Assess the importance of features using ensemble methods like Random Forests or Gradient Boosting

What is next?

How to Reduce the dimensionality of the feature space



Understand the Problem and Data

Perform Exploratory Data Analysis (EDA)

Feature Engineering

Feature Selection

Dimensionality Reduction

Evaluate Feature Importance

Iterate and Experiment

Domain Knowledge Integration

Validate and Test

Master in Artificial Intelligence

Feature Engineering III